УДК 519.8

И.В. Козин
Запорожский национальный университет МОН Украины, Украина
Украина, 69600, г. Запорожье, ул. Жуковского, 66

ЭВОЛЮЦИОННЫЙ АЛГОРИТМ ОПТИМАЛЬНОЙ КЛАССИФИКАЦИИ

I.V. Kozin
Zaporizhzhya National University MES of Ukraine, Ukraine
Ukraine, 69600, c. Zaporizhzhya, Zhukovsky str., 66

EVOLUTIONARY ALGORITHM OF THE OPTIMAL CLASSIFICATION

І.В. Козін
Запорізький національний університет МОН України, Україна
Україна, 69600, м. Запоріжжя, вул. Жуковського, 66

ЕВОЛЮЦІЙНИЙ АЛГОРИТМ ОПТИМАЛЬНОЇ КЛАСИФІКАЦІЇ

В работе представлены результаты исследования эволюционно-фрагментарной модели для поиска оптимальной классификации. Показано, что задача поиска оптимальной классификации на конечном множестве может рассматриваться как задача на фрагментарной структуре. Предложена эволюционно-фрагментарная модель задачи классификации на множестве перестановок.
Ключевые слова: задача классификации, фрагментарная структура, эволюционная модель.

Research results evolutionarily fragmented model to find the optimal classification presented in the work. It has shown that the problem of finding an optimal classification on a finite set considered as a challenge for the fragmented structure. An evolutionarily fragmented model of the problem of classification on a set of permutations is proposed.
Key words: classification problem, fragmented structure, evolutionary model.

У роботі представлено результати дослідження еволюційно-фрагментарної моделі для пошуку оптимальної класифікації. Показано, що задача пошуку оптимальної класифікації на кінцевій множині може розглядатися як задача на фрагментарній структурі. Запропоновано еволюційно-фрагментарну модель задачі класифікації на множині перестановок.
Ключові слова: задача класифікації, фрагментарна структура, еволюційна модель.

Введение

Одной из основных задач исследования операций является задача классификации. Этой задаче посвящены многочисленные публикации[1-3]. Известны много подходов к поиску решения задачи классификации на конечных множествах. Тем не менее, говорить сегодня о том, что эта задача исследована полностью еще рано. В этой статье мы не будем обсуждать принципы выбора классификации, которые относятся большей частью к области теории принятия решений. Речь пойдет лишь об одной постановке этой задачи, основанной на минимизации диаметров кластеров – элементов классификации [2]. Как известно, эта задача относится к числу *NP* – полных задач [4]. Точный алгоритм полиномиальной трудоемкости для поиска оптимальной классификации неизвестен.

Для поиска решения этой задачи часто используются эвристические алгоритмы [5]. В работе предлагается метод, основанный на комбинации

эволюционного и фрагментарного алгоритмов. Подобная технология показала хорошие результаты для ряда сложных дискретных оптимизационных задач [6-7].

Постановка задачи. Рассмотрим задачу классификации в следующей постановке: задано конечное множество объектов $X = \{x_1, x_2, ..., x_N)$. Определена мера различия между объектами. Эта мера задается матрицей $D = \{d_{ij}\}_{i,j=1}^{N}$. Причем $d_{ij} \geq 0, \quad d_{ij} = d_{ji}, \quad d_{ii} = 0, \quad i, j = 1,2,...,N$. Каждый элемент матрицы d_{ij} определяет различие между соответствующими объектами. Задача классификации состоит в разбиении множества объектов на k непустых попарно непересекающихся классов

$$X = X_1 \cup X_2 \cup ... X_k, \quad X_i \cap X_j = \varnothing, \quad X_i \neq \varnothing \quad \forall i \neq j, \; i, i = 1,2,....N.$$

Диаметром класса X_s, будем называть число $\delta_s = \max\limits_{x_i, x_j \in X_s} d_{ij}, \quad s = 1,2,...,k$.

Оптимальной классификацией называется такое разбиение множества X, при котором величина $\Delta = \max\limits_{s=1,2,...,k} \delta_s$ минимальна. Величину Δ будем называть диаметром разбиения.

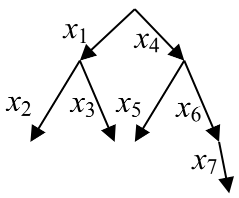
В [4] доказано, что задача классификации в приведенной выше постановке является NP-полной.

Фрагментарная структура

В соответствии с [8] упорядоченной фрагментарной структурой (X, E) на конечном множестве X называется семейство его упорядоченных подмножеств $E = \{E_1, E_2, ..., E_n\}$ такое, что

$$\forall E_i \in E, \quad E_i \neq \varnothing, \quad E_i = \{x_1, x_2, ..., x_{k_i}\}, \quad \forall s \leq k_i, \{x_1, x_2, ..., x_s\} \in E.$$

Простым, но достаточно общим примером фрагментарной структуры могут служить ориентированные пути в ориентированном корневом дереве. На рис.1 показано ориентированное дерево с семью дугами и наборы путей, образующих упорядоченную фрагментарную структуру



$$E_1 = \varnothing; \quad E_2 = \{x_1\}; \quad E_3 = \{x_4\};$$
$$E_4 = \{x_1, x_2\}; \quad E_5 = \{x_1, x_3\}; \quad E_6 = \{x_4, x_5\};$$
$$E_7 = \{x_4, x_6\}; \quad E_8 = \{x_4, x_6, x_7\};$$

Рис.1 Ориентированное дерево и соответствующая ему фрагментарная структура

Элементы множества E называются допустимыми фрагментами. Одноэлементные множества, которые являются допустимыми фрагментами, будем называть элементарными фрагментами. Фрагмент называется максимальным, если он не является подмножеством никакого другого фрагмента. Всякий допустимый фрагмент можно построить из пустого множества, последовательно добавляя к нему элементы так, чтобы на каждом шаге такой процедуры полученное подмножество было допустимым фрагментом.

Максимальный фрагмент может быть построен с помощью фрагментарного алгоритма:

а) элементы множества X линейно упорядочиваются;

б) на начальном шаге выбирается пустое множество $X_0 = \varnothing$;

в) на шаге с номером k+1 выбирается первый по порядку элемент $x \in X \setminus X_k$, такой, что $X_k \bigcup \{x\} \in E$;

г) алгоритм заканчивает работу, если на очередном шаге не удалось найти элемент $x \in X \setminus X_k$ с требуемым свойством.

Результат применения фрагментарного алгоритма определяется заданным линейным порядком на множестве *X*. Таким образом, любой максимальный фрагмент может быть описан некоторой перестановкой элементов множества *X*.

Пусть теперь каждому максимальному фрагменту приписано неотрицательное число – вес фрагмента. Тогда возникает естественное отображение множества перестановок во множество весов, которое каждой перестановке $s \in S_n$ ставит в соответствие неотрицательное число – вес максимального фрагмента, который определяется перестановкой *s* элементов базового множества *X*.

Покажем, что задача классификации на конечном множестве *X* объектов в приведенной выше постановке может быть сведена к задаче поиска максимального фрагмента упорядоченной фрагментарной структуры с минимальным весом. Элементарными фрагментами этой структуры будут все одноэлементные подмножества множества *X*.

Упорядочим элементы множества *X* в определенном порядке $(x_1, x_2, ..., x_N)$. С каждым таким упорядочением связано C_{N-1}^{k-1} различных разбиений множества *X* на *k* непустых попарно непересекающихся подмножеств. Каждое такое разбиение *Q* определяется выбором *k*-1 границ из *N*-1 возможных между элементами множества *X*, выстроенных в заданном порядке:

$$x_1, x_2, ..., x_{s_1} \mid x_{s_1+1}, x_{s_1+2}, ..., x_{s_2} \mid .... \mid x_{s_{k-1}+1}, x_{s_{k-1}+2}, ..., x_N .$$

Построение каждого такого разбиения осуществляется алгоритмом константной трудоемкости. Поставим в соответствие каждому разбиению *Q* его диаметр Δ_Q. Из всех разбиений, соответствующих заданному порядку на множестве *X* выберем то, диаметр которого минимален (если таких разбиений несколько, то выбираем одно из них по детерминированному правилу). Таким образом, с каждой перестановкой элементов связана некоторая классификация элементов множества *X*. При заданной матрице $D = \{d_{ij}\}_{i,j=1}^{N}$ различий между элементами множества *X* трудоемкость алгоритма отыскания этой классификации есть $O(N^{k+1})$. Максимальным фрагментом в рассматриваемом случае является любое линейное упорядочение множества *X*. Число таких упорядочений есть *N*! и потому задача перебора всех фрагментов является трудной. Для быстрого поиска классификаций, которые близки к оптимальной, предлагается следующая стандартная эволюционная модель на фрагментарной структуре.

Эволюционная модель

Основные составляющие эволюционной модели следующие [9-11]:

- базовое множество решений – множество допустимых решений *V*, на котором ищется оптимальное решение задачи;

- оператор построения начальной популяции: процедура, которая позволяет выделить на множестве всех допустимых решений его подмножество $W \subseteq V$ для последующей эволюции;

- критерий селекции – алгоритм вычисления значения по целевой функции по заданной перестановке на элементах базового множества решений, который позволяет упорядочивать решения в рамках заданной популяции;

- оператор кроссовера $K : V \times V \to V$, позволяющий по двум допустимым решениям-родителям построить новое решение-потомок из множества допустимых решений;

- оператор мутации $M : V \to V$;

- оператор отбора, который выделяет множество пар в W для выполнения операции кроссовера;

- оператор эволюции, позволяющий строить новые популяции из множества родителей и потомков;

- правило остановки, которое определяет условие остановки эволюционного алгоритма.

Опишем кратко принцип работы эволюционного алгоритма. На начальном шаге с помощью оператора начальной популяции строится множество решений W_0. На каждом очередном шаге предполагается заданным некоторое множество допустимых решений - текущая популяция. На первом шаге это множество $W = W_0$. Для каждого из элементов множества W вычисляется значение целевой функции. Далее с помощью оператора отбора в текущей популяции W выбирается множество пар для кроссовера. К каждой паре из выбранного множества пар применяется оператор кроссовера, а затем к результату кроссовера применяется оператор мутации. Таким путем находится множество элементов – потомков $\tilde{W}$.

К промежуточной популяции $W \cup \tilde{W}$, которая является объединением текущей популяции и множества потомков, применяется оператор эволюции, который выделяет на этом множестве новую текущую популяцию. Процесс эволюции повторяется до тех пор, пока не будет выполнено условие остановки эволюционного алгоритма.

Эволюционно-фрагментарный алгоритм (ЭВФ-алгоритм) является комбинацией эволюционного и фрагментарного алгоритма. Опишем соответствующую эволюционную модель и принцип действия такого алгоритма.

В качестве множества допустимых решений рассматривается подмножество максимальных фрагментов на заданной упорядоченной фрагментарной структуре. Каждый фрагмент из этого множества определяется соответствующим линейным упорядочением элементов множества X. Таким образом, любому допустимому решению соответствует определенная перестановка чисел 1,2,…, N, где N-количество элементарных фрагментов. Для максимального фрагмента определено значение критерия селекции.

Базовое множество V эволюционной модели – это множество $S_N = \{i_1, i_2, ..., i_N\}$ всех перестановок чисел 1,2,…,N. Оператор построения начальной популяции выделяет подмножество заданной мощности Q из множества V.

Критерий селекции устроен следующим образом: по заданной перестановке фрагментов с помощью фрагментарного алгоритма строится максимальный допустимый фрагмент. Вычисляется значение целевой функции задачи для этого фрагмента.

Опишем теперь оператор кроссовера. Пусть $u=(u_1,u_2,...,u_N)$ и $v=(v_1,v_2,...,v_N)$ – две произвольные перестановки. Перестановка - потомок строится следующим образом: последовательности u и v просматриваются с начала. На m-м шаге выбирается наименьший из первых элементов последовательностей и добавляется в новую перестановку - потомок. Затем этот элемент удаляется из двух последовательностей-родителей. Например,

$$K((3,2,8,6,5,1,4,7),(4,7,2,5,1,6,8,3)) = (3,2,4,7,5,1,6,8)$$

Оператор мутации M выполняет случайную транспозицию (замену местами двух элементов) в перестановке.

Оператор селекции выбирает случайным образом набор пар из заданного числа пар во множестве перестановок текущей популяции.

Оператор эволюции элементы промежуточной популяции упорядочивает в последовательность по убыванию значения критерия селекции. В качестве новой текущей популяции выбираются первые Q элементов последовательности.

Обычное правило остановки - количество поколений достигло предельной границы L. Лучшая по значению критерия селекции перестановка из последней построенной популяции определяет приближенное решение задачи.

Предложенный подход является универсальным и позволяет применять один и тот же эволюционный алгоритм к любой оптимизационной задаче на конечной фрагментарной структуре.

Результаты тестирования

Рассмотренный выше метод был применен к задаче разбиения множества точек на плоскости на заданное количество классов. Мерой близости точек являлось обычное эвклидово расстояние. Входными параметрами при описании серии случайных задач являются: число точек плоскости N; диапазоны изменения координат $[0,A],[0,B]$.

С помощью генератора случайных чисел генерируется наборы точек $(x_1,x_2,...,x_N),\quad x_i=(a_i,b_i),\quad a_i\in[0,A],b_i\in[0,B],\quad i=1,2,...,N$. Число точек N выбиралось в промежутке 300-1000, число классов -10. Параметры эволюционного алгоритма: Размер популяции -1000, количество поколений-500, 30 пар для селекции в каждом поколении, вероятность мутации – 0,05.

Для поиска решений использовалась авторская программа, реализующая эволюционную модель на перестановках. Программа написана на VBA, под управлением WINDOWS XP/7. Вычисления проводились на компьютере ACPIx64-based PC с процессором 4x3400 MHz. Время расчета для каждой индивидуальной задачи было ограничено одной минутой.

Сравнивались решения, полученные иерархическими агломеративными методами [12] с решениями, полученным путем применения ЭВФ алгоритма. В серии из 1000 случайно сформированных задач в 98% случаев лучшей оказывалась классификация, полученная с помощью ЭВФ алгоритма.

Выводы

Теоретические результаты и результаты численных экспериментов показывают, что ЭВФ – алгоритм может достаточно эффективно использоваться как эвристический алгоритм при решении задачи классификации. ЭВФ-алгоритм является управляемым, качество его работы может возрастать при изменении ряда параметров алгоритма, таких как величина популяции, число пар для селекции, количество поколений, количество эволюций и т.д.

Литература

1. Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: Классификация и снижение размерности. Справочное издание / Под ред. С.А. Айвазяна. – М.:Финансы и статистика, 1989. – 607 с.
2. В.Перепелица Задачи классификации и формирование знаний / В.Перепелица, И. Козин, Э.Терещенко //Lap LAMBERT Academic Publishing, Germany -2012, 196p.
3. Кендалл М.Дж., Стьюарт А. Многомерный статистический анализ и временные ряды. - М.: Наука, 1976. – 736 с.
4. Brucker, P.,On the complexity of clustering problems. In: Beckmenn, M., Kunzi, H.P. (Eds.), Optimisation and Operations Research. Lecture Notes in Economics and Mathematical Systems. Springer, Berlin, vol. 157, 1978, pp. 45-54.
5. Genetic clustering for automatic evolution of clusters and application to image classification. Pattern Recognition. v35., 2002, 1197-1208.
6. Козин И.В. Фрагментарные алгоритмы в системах поддержки принятия решений /И.В.Козин// Питання прикладної математики і математичного моделювання, збірник наукових праць. ДНУ: Дніпропетровськ, 2006. — С. 131—137.
7. Козин И.В. Фрагментарный алгоритм для задачи симметричного размещения /И.В.Козин// Радиоэлектроника, информатика, управление.2005, №1. — С. 76—83.
8. Козин И.В.О свойствах фрагментарных структур/И. В.Козин, С.И.Полюга/ Вісник Запорізького національного університету. Математичне моделювання і прикладна механіка. – 2012. – № 1. – С. 99-106.
9. Holland J. H. Adaptation in Natural and Artificial Systems / J. H. Holland. – Boston, MA : MIT Press. –1992. -288p.
10. Курейчик В. М. Генетические алгоритмы. Состояние. Проблемы. Перспективы / В. М. Курейчик // Известия РАН. ТиСУ. - 1999. - №1. - С. 144-160.
11. Скобцов Ю.А. Основы эволюционных вычислений: учеб.пособ./Ю.А.Скобцов - Донецк: [ДонНТУ], 2008. - 326 с.
12. Ким Дж. О. Факторный, дискриминантный и кластерный анализ / Дж. О. Ким, Ч. У. Мюллер, У. Р. Клекка и др. М. : Финансы и статистика, 1989. - 215 с.

Literatura

1. Ayvazyan S.A., Buhshtaber V.M., Enyukov I.S., Meshalkin L.D. Prikladnaya statistika: Klassifikaciya i snizhenie razmernosti. Spravochnoe izdanie / Pod red. S.A. Ayvazyana. - M.:Finansyi i statistika, 1989. - 607 s.
2. V.Perepelica Zadachi klassifikacii i formirovanie znaniy / V.Perepelica, I. Kozin, YE.Teres`henko //Lap LAMBERT Academic Publishing, Germany -2012, 196p.
3. Kendall M.Dzh., St`yuart A. Mnogomernyiy statisticheskiy analiz i vremennyie ryadyi. - M.: Nauka, 1976. - 736 s.
4. Brucker, P.,On the complexity of clustering problems. In: Beckmenn, M., Kunzi, H.P. (Eds.), Optimisation and Operations Research. Lecture Notes in Economics and Mathematical Systems. Springer, Berlin, vol. 157, 1978, pp. 45-54.
5. Genetic clustering for automatic evolution of clusters and application to image classification. Pattern Recognition. v35., 2002, 1197-1208.
6. Kozin I.V. Fragmentarnyie algoritmyi v sistemah podderzhki prinyatiya resheniy /I.V.Kozin// Pitannya prikladnoy matematiki i matematichnogo modelyuvannya, zbirnik naukovih prac. DNU: Dnipropetrovsk, 2006. - S. 131-137.
7. Kozin I.V. Fragmentarnyiy algoritm dlya zadachi simmetrichnogo razmescheniya /I.V.Kozin// Radioyelektronika, informatika, upravlenie.2005, №1. - S. 76-83.
8. Kozin I.V., O svoystvah fragmentarnyih struktur/ I. V.Kozin, S.I.Polyuga // Visnik Zaporizkogo nacionalnogo universitetu. Matematichne modelyuvannya i prikladna mehanika. - 2012. - № 1. - S. 99-106.
9. Holland J. H. Adaptation in Natural and Artificial Systems / J. H. Holland. – Boston, MA : MIT Press. –1992. -288p.
10. Kureychik V. M. Geneticheskie algoritmyi. Sostoyanie. Problemyi. Perspektivyi / V. M. Kureychik // Izvestiya RAN. TiSU. - 1999. - №1. - S. 144-160.
11. Skobcov YU.A. Osnovyi yevolyucionnyih vyichisleniy : ucheb. posob./YU.A.Skobcov - Doneck: [DonNTU], 2008. - 326 s.
12. Kim Dzh. O. Faktornyiy, diskriminantnyiy i klasternyiy analiz / Dzh. O. Kim, CH. U. Myuller, U. R. Klekka i dr. M. : Finansyi i statistika, 1989. - 215 s.

RESUME

I.V. Kozin

Evolutionary algorithm of the optimal classification

The article discusses a known problem - the problem of optimal classification. This problem occurs in many applications in engineering and economics. It is known that in most cases the classification problem on a finite set is NP-hard. For such problems today are not known polynomial complexity algorithms for finding solutions. Therefore, in an optimal classification is justified use of metaheuristics. The initial data of the problem given the proximity matrix between elements of the set, which determines the degree of similarity between these elements. The number of classes is considered to be specified. The criterion for classification is the greatest quality of the diameters of classes. To find the approximate optimal classification is proposed to use evolutionary and fragmented model. This model has worked well in finding optimal solutions to many difficult problems of discrete optimization.

The article introduces the concept of detail-oriented fragmented structure. It describes fragmented algorithm that allows for a given order of elementary fragments to build the maximum allowable fragment of a fragmented structure.

It is shown that the problem of optimal classification may be presented as a problem of optimization on a fragmented structure. This reduces the problem to the optimization problem of finding an optimal solution of a combinatorial problem on the set of permutations. Evolutionary algorithm is proposed, for which the base set - the set of permutations of elementary fragments. Numerical experiments on a large number of test problems showed the effectiveness of the proposed approach for finding approximate solutions of the problem of optimal classification.

І.В. Козін

Еволюційний алгоритм оптимальної класифікації

У статті розглянуто відому проблему – задачу оптимальної класифікації. Ця задача зустрічається в численних застосуваннях у техніці та економіці. Відомо, що в більшості випадків задача класифікації на скінченній множині є NP-важкою. Для таких задач на сьогоднішній день невідомі алгоритми поліноміальної трудомісткості для пошуку рішення. Тому в задачах оптимальної класифікації виправдане застосування метаевристик. Як вихідні дані задачі, задано матрицю близькості між елементами множини, яка визначає ступінь схожості цих елементів. Кількість класів вважається заданим. Критерієм якості класифікації є найбільший з діаметрів класів. Для пошуку наближеної оптимальної класифікації запропоновано використовувати еволюційно-фрагментарну модель. Ця модель добре зарекомендувала себе при пошуку оптимальних рішень багатьох важких задач дискретної оптимізації.

У статті докладно представлено концепцію орієнтованої фрагментарної структури. Наведено фрагментарний алгоритм, який дозволяє для заданого порядку елементарних фрагментів побудувати максимально допустимий фрагмент фрагменттарної структури. Показано, що задача оптимальної класифікація може бути представлена як задача оптимізації на фрагментарної структури. Це дозволяє звести задачу оптимі-зації до задачі пошуку оптимального рішення комбінаторної задачі на множині перестановок. Запропоновано еволюційний алгоритм, для якого базова множина – множина перестановок елементарних фрагментів. Чисельний експеримент на великій кількості тестових задач показав ефективність пропонованого підходу для пошуку наближеного рішення задачі оптимального класифікації.

Поступила в редакцию 25.08.2015

